**FREQUENT PATTERNS AND ASSOCIATION RULES**

**PROJECT 1 REPORT**

on

**Prediction of fake post and the type of post that tend to be viral on 2016 US elections on Social Media website Facebook**

by

Shubharaj Arsekar

Ravi Teja Neela

Texas A&M University-Corpus Christi

Major: Computer Science

Under the supervision of Dr. Longzhuang Li

Summer 2017

Table of Contents

[Introduction: 3](#_Toc485944732)

[Data Mining 3](#_Toc485944733)

[Data Mining Process 3](#_Toc485944734)

[Mining Tool Description: 4](#_Toc485944735)

[Objectives: 4](#_Toc485944736)

[Potential Benefits: 4](#_Toc485944737)

[Dataset Description 5](#_Toc485944738)

[Title: 5](#_Toc485944739)

[Year: 5](#_Toc485944740)

[Source Information: 5](#_Toc485944741)

[Data Set Information: 5](#_Toc485944742)

[Context 5](#_Toc485944743)

[Content 5](#_Toc485944744)

[Number of instances 5](#_Toc485944745)

[Number of attributes 5](#_Toc485944746)

[Process Design 7](#_Toc485944747)

[Pre-processing Data 7](#_Toc485944748)

[Association Rule 9](#_Toc485944749)

[Fake Data 9](#_Toc485944750)

[Sharing and Reaction 10](#_Toc485944751)

[Rules 10](#_Toc485944752)

[Recommendation 11](#_Toc485944753)

# Introduction:

## Data Mining

Data mining is the process of sorting of large data sets to recognize patterns and obtain connections, structures, and relationships between the data in the dataset providing a solution in the form of Data Analysis. Data mining allows to predict sensitive trends which may be used to provide a better solution to the recognized problem. An example for the same would be based on the share market data, the data generated in share market has large volumes and is produced on an hourly rate on daily basis. Data miner use specific tools build for data mining to find frequent patterns and relationships and can predict trends and rules to improve the profit of a share values based on the previous data obtained, such rules are valuable for a company by preventing the losses that may occur due to market crash.

## Data Mining Process

Data mining is a five-step process in general:

1.) Organization collects data and dumps the raw data into their warehouse. This data is not processed and there is no relationship between the data. The data may be cross domain.

2.) They store and manage the data on cloud or in-house server having redundancy and backup.

3.) Data Analyst, Management teams, Data mining experts and Information technology sort the data based on client and company requirement. This is basically the pre-processing where the data is filtered as required to obtain meaningful pattern.

4.) The above step is performed both manually and using the data mining tools. There is various operator like replacement of missing value etc. This step is where the processes and algorithm are applied on processed, filtered and extracted data.

5.) The last step is to pull the rules and patterns by analysing the results obtain to draw meaningful prediction that can be used further for the benefits or as a training set for the incoming data.

The results and recommendations obtained from these patterns, structure and relationship is discovered by using two implementations - predictive and descriptive.

The descriptive is implemented on the datasets below because the main idea is to find the human readable patterns which can give useful recommendation for future works on the same domain data. This implementation has several techniques/rules that could be chosen which are Classification, Segmentation/Clustering, Association, etc.

## Mining Tool Description:

RapidMiner formerly known as YALE (Yet Another Learning Environment), is a data science software platform developed on an open core model by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization

# Objectives:

Domain: Social Media Data

Data Set source: Kaggle.com

Data Source: fact-checking-facebook-politics-pages.csv

The objective of this project is to generate association rules and generate frequent patterns using the association algorithm-Apriori, and provide recommendation and future trends to notify the issues to find the fake data posts and find the patterns in the how viral the post may become based on the type of posts (video, link, text and photos).

## Potential Benefits:

The Association rule is implemented because the higher frequency of human readable patterns. In the selected data set, we can predict the correlation between the trends, partisan among the voting banks, and how much the fake data can affect and swing the election trends that may happen in the next coming election after applying the above generated rules.

# Dataset Description

Title: Fact checking Facebook politics pages

Year: 2016

Source Information: https://www.kaggle.com/buzzfeed/fact-checking-facebook-politics-pages

## Data Set Information:

### Context

During the 2016 US presidential election, the phrase “fake news” found its way to the forefront in news articles, tweets, and fiery online debates the world over after misleading and untrue stories proliferated rapidly.

[BuzzFeed News](https://www.kaggle.com/buzzfeed) analyzed over 1,000 stories from hyperpartisan political Facebook pages selected from the right(conservative), left(liberal), and mainstream(neutral) media to determine the nature and popularity of false or misleading information they shared.

### Content

This dataset supports the original story [“Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate”](https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis?utm_term=.kq9kqJDZ2) published October 20th, 2016.

Number of instances: 2283 rows

Number of attributes: 12

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Datatype** | **Values** |

|  |  |  |  |
| --- | --- | --- | --- |
| Account\_Id | Facebook page account ID | Numeric | Continuous |
| Post\_Id | POST\_ID | Numeric | Continuous |
| Category | Type of Source (Left, Right & Mainstream) | String | Left, Right & Mainstream |
| Page | Name of the page | String | Continuous |
| Post\_URL | URL where the post was found | String | Continuous |
| Date Published | Date the post was published | Datetime | Dec 21 2016 |
| Post\_Type | Type of post shared (Video or Link) | String | Video, Link, Text and Photo |
| Rating | “Truth” rating made by Buzz feed | String | Mostly True, Mostly False, Mixture of true and false, No factual Content |
| Debate | Video showing a debate | String | YES or NO |
| Share\_Count | Number of Shares | Numeric | Continuous |
| Reaction\_Count | Number of Reactions (Likes) | Numeric | Continuous |
| Comment\_Count | Number of Comments | Numeric | Continuous |

Table 1: Attribute Description

Below is a sample post over Facebook which is having 68 likes and 2 shares with lots of comments on it.



Fig 1: Example Post from the data set URL

The following steps have been traversed for mining and then analysing to obtain fruitful rules using Rapid Miner Tool

# Process Design

A screenshot of a computer

Description generated with very high confidence

Fig 2: Overview of the process design

Pre-processing Data: This is divided into 4 parts:

* + The data set contains missing values in the share\_count, comment\_count and reaction and the data type is numeric and continuous so we use the average values to replace the missing value.
  + The other attribute that has the missing data is Debate which a nominal data. The data displays a “yes” when there is a video post and the remain post have the value as missing, so we replace the value with a “no” assuming the video is the only one where there is debate happening.

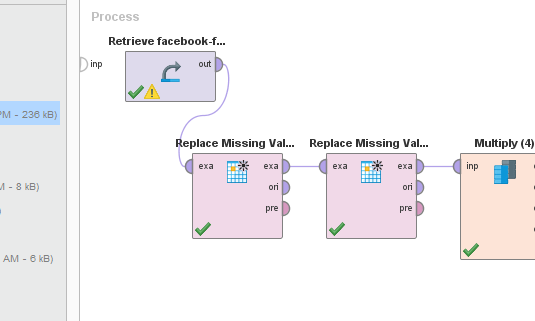


Fig 3: Replace Missing Value

* + The attributes category (left, right and mainstream), rating (mostly true,mostly false, mixture of true and false, no factual content), post\_type(video, photos, links, text) are polynomial (nominal) having more than two instances and are converted to binomial (true and false) so the data is split into multiple columns to maintain the data consistency.

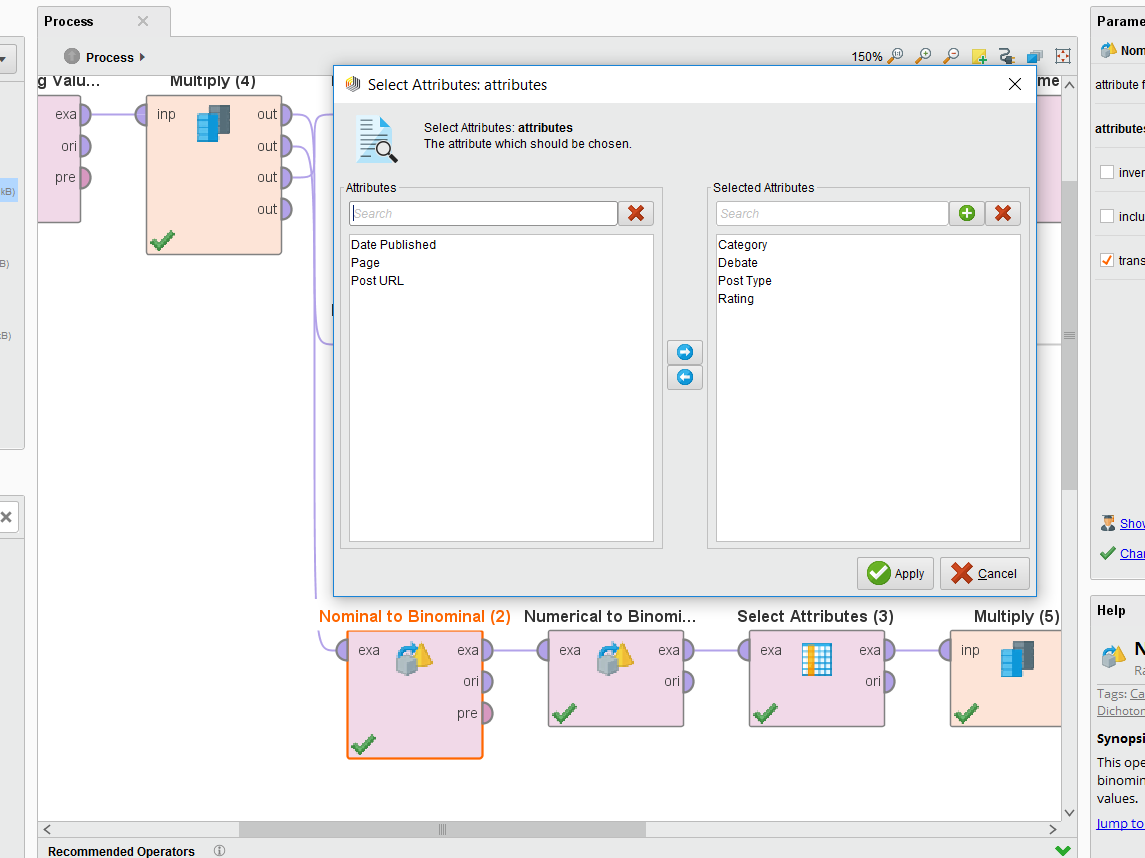
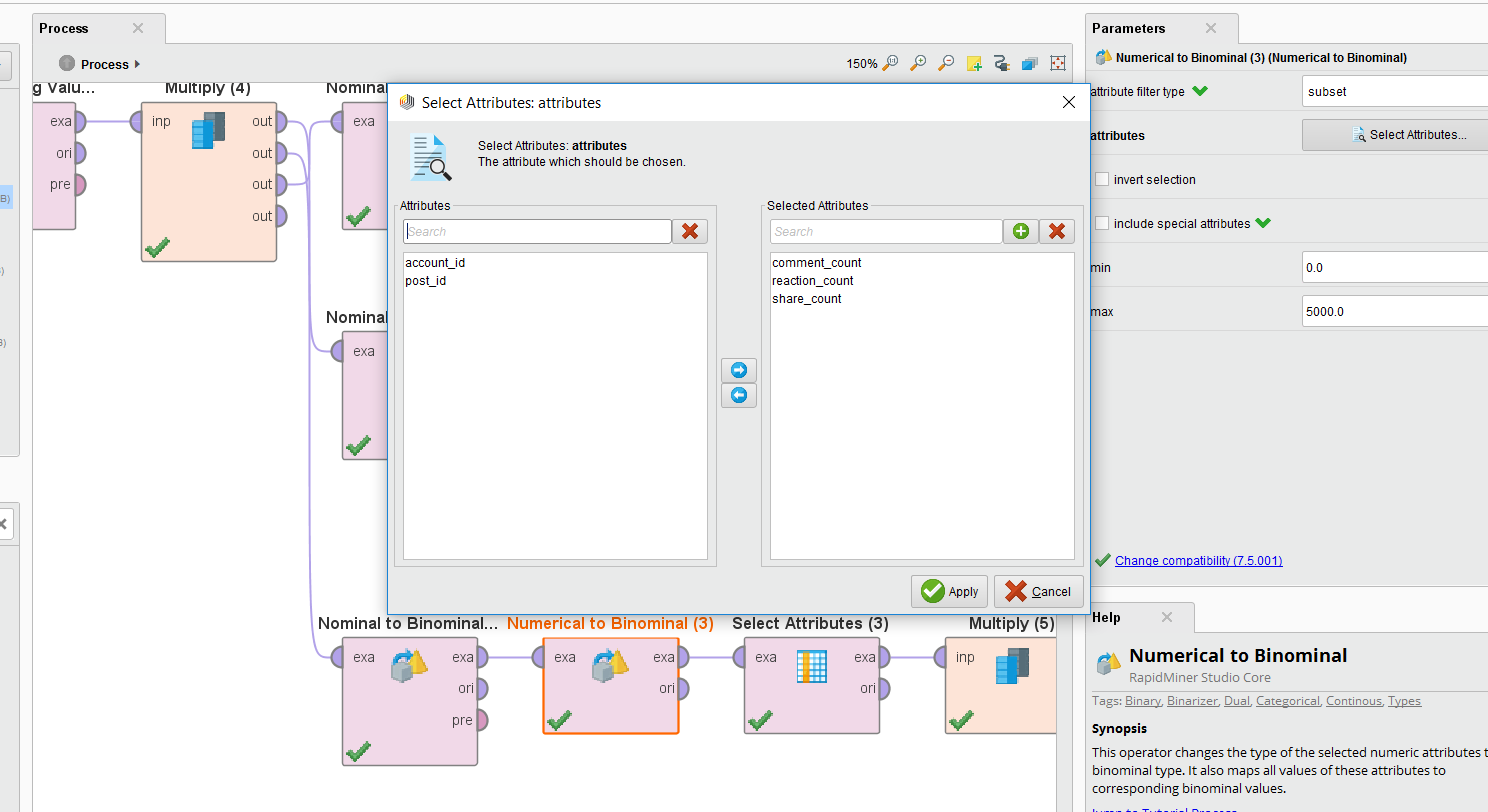


Fig 4: Polynomial to Binomial

* + The numerical attributes share\_count, comment\_count, and reaction\_count are converted to binomial based on values above average are considered to be true (highly popular-tend to be viral)

Fig 5: Numerical to Binomial

Association Rule: The apriori algorithm is then applied to generate descriptive rules

### Fake Data

* If Ratings = Mostly True = False 440 ==> Category = Right=True 347 confidence:(0.79)
* If Rating = Mostly True = False 613 ==> Category = Mainstream=False 553 confidence:(0.9)
* If Rating = Mostly True = False 340==> Category = Left=False 553 confidence:(0.9)

🡪 **When the post is false we found that left and Mainstream posted fake data having a confidence of 0.9**

### Sharing and Reaction

**Filters: Category = left; Rating = Mostly True;**

* If Category = Left=True, Post-type = Link=True and Rating = Mostly True=True

==> Share-Count = True 227 ==> Reaction-Count = True 220 confidence:(1)

* If Category = Left=True, Post-type = Video=True and Rating = Mostly True=True

==> Share-Count= True 192 ==> Reaction-Count=true 190 confidence: (0.99)

* If Category = Left=True, Post-type = Photo=True and Ratings = Mostly True = True

==> Share-Count= False 215 ==> Reaction-Count=true 209 confidence: (0.97)

* If Category = Left=True, Post-type = Text=True and Ratings = Mostly True = True

==> Share-Count =True 206 ==> Reaction-Count =True 202 confidence: (0.98)

**If the Post type was link or video there were more chances of it being shared and therefore higher reaction to the post.**

Rules: The filter is added to get meaningful examples for specific rules.

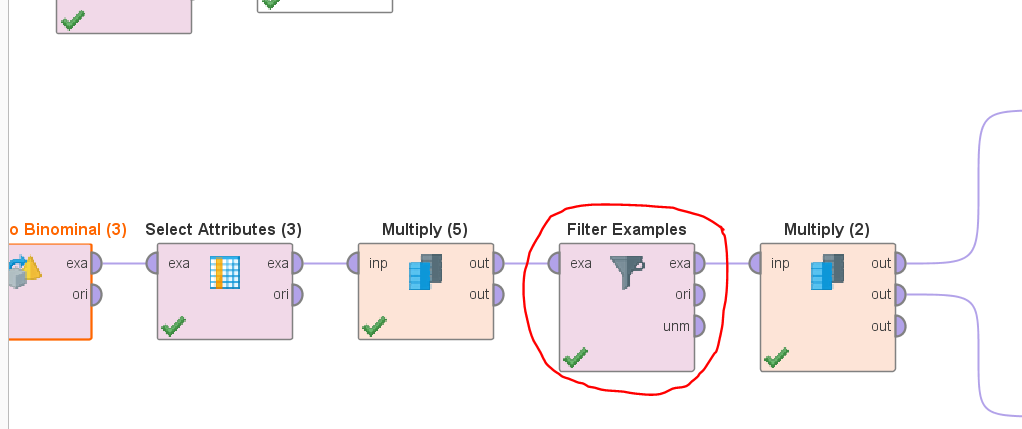
**

Fig 6: Filter Example

# Recommendation

* + The rules provide how much fake content is released to obtain monopoly in the elections. The effect of the data post recommends on which side the elections are swinging towards the power distribution.
  + This also can be used by the election moderator to scrutinize the political aspects.
  + The post type also plays a viral role on the trends. This rule can be implemented in the general social media data. It was found that if the post types were video or link then they tend to view more by more number of shares and thereby had higher reactions(like) and greater comments counts.